Елена Паскалева
(София)

ФОРМАЛИЗАЦИЯ НА РЕЧНИКОВАТА СТАТИЯ И ПРОИЗВОДСТВОТО НА РЕЧНИЦИ – ОПИТИ, ЖЕЛАНИЯ, ВЪЗМОЖНОСТИ

Abstract: The problems of standardization and formalization of the structure of dictionary entry are discussed in their resolving within the frames of several European projects (with Bulgarian participation). A software system with rich interface facilities is reported. It performs the automatic conversion from printed dictionary data to database presentation. The resulting presentation of large Bulgarian explanatory dictionary in html format is installed in a dictionary server in intranet implementation. The server is out of use for license reasons but the database is open for discussions and investigations.

Key words: lexical data bases, dictionary standards, intranet server, e-dictionary

Речниковата статия като обект на формализация

Интересът, теоретическото позициониране, експериментите, успехите и приложенията, свързани със задачата да се формализира описанието на речниковата статия, са породени от многопосочния натиск на различни фактори в езиковите технологии. Първо – повишеното внимание към многоезиковите широкообемни построения, чиято най-чиста еманация на равнище Дума се явява речникът. Второ – необходимата стандартизация на разнообразните лингвистични обекти; трето – повишените изисквания на речниковото производство, обслужващо комуникационните нужди на многоезичието. Неслучайно програмната статия *Encoding Dictionaries* е поместена през 1995 г. в броя на *Computers and the Humanities*, посветен на стандарта Text Encoding Initiative в широкия диапазон на възможните му приложения (Ide and Veronis 1995: 167–195).

Първи опити за формализация и стандартизация на речниковата статия в многоезиков аспект

В рамките на проекта GLOSSER (Copernicus 343) в далечния предприсъединителен период (1993–1995) една платформа за изучаване на английски език, разработена за естонски, фински, унгарски и български, ползва освен подравнени двуезикови корпуси речникова информация от общ за всички езикови двойки речник, известната серия от т.нар. полудвуезични речници на Кернерман (KSBD). На базата на тълковния *Chambers Concise Usage*

Dictionary всяка една от неговите 30 англо–Х версии задава допълнително в речниковата статия преводите на речниковата единица, нейното тълкуване, производни и примери. Тази му унифицирана структура позволи речникът да бъде използван в многоезикови приложения като цитираното в GLOSSER. Още повече, че унифицираното съдържание на статията в този речник се допълва и от унифициран стандарт на представяне – родителя на всички стандарти за представяне на документи: SGML – Standard Generalized Markup Language (вж. фиг. 1). Проблеми на авторското право ограничиха реалното запълване на този общ стандарт само до илюстративните сегменти на системата (Nerbonne et al. 1997: 135–138).

```
<art>
<en>boat</EN>
<PRON>[bout]</PRON>
<POS>noun</POS>
<SEM>
<SN>1</SN>
<DEF>a small vessel for travelling over water</DEF>
<EX>We'll cross the stream by boat.</EX>
<BG>лодка</BG>
<SN>2</SN>
<DEF>a larger vessel for the same purpose; a ship</DEF>
<EX>to cross the Atlantic in a passenger boat.</EX>
<BG>кораб</BG>
<SN>3</SN>
<DEF>a serving-dish shaped like a boat</DEF>
<EX>a gravy-boat.</EX>
<BG>сосиера</BG>
<GRH>
<POS>verb</POS>
<DEF>to sail about in a small boat for pleasure</DEF>
<EX>They are boating on the river.</EX>
<BG>возя се на лодка</BG>
<ENS>`boatman</ENS>
<POS>noun</POS>
<DEF>a man in charge of a small boat in which fare-paying passengers are carried</DEF>
<BG>лодкар</BG>
<ENS>in the same boat</ENS>
<DEF>in the same, usually difficult, position or circumstances</DEF>
<EX>We're all in the same boat as far as low wages are concerned.</EX>
<BG>на едно и също положение
</BG>
</SEM>
</ART>
```

Фиг. 1. Речниковата статия на *boat* в англо-българската версия на KSBD

Подходи към прехода от печатен към електронен речник

Както във всички сектори на компютърната лингвистика, където моделирането на езиков обект е свързано с ползването на реални компютърни приложения, подходите към компютризацията на обекта са двупосочни – от общото към частното (top down) или обратно (bottom up). В нашия случай двата отправни пункта са: структурирано формално представяне на статията в различни речници, за да се стигне до обобщената формална схема или едно генерално формално представяне, които да е запълнено в отделните случаи с реални лингвистични обекти. Гореописаният опит за общо моделиране на

речниковата статия в многоезиково изпълнение е илюстрация на първия подход, улеснен от унифицираното представяне на отделните речници в концепцията на Кернерман.

Моделиране на речника в обратната посока – от генералната схема към конкретната реална, срещаме в проекта CONCEDE, програма INCO-COPERNICUS (1998–2000). Участващите в проекта шест източноевропейски страни на базата на паралелни текстове и налични речници (в различна степен на електронизация) поставят основите на генерална речникова схема чрез създаването на речниковия документен стандарт DTD (Document Type Definition), който и до днес се използва за описание и стандартизация на речникарска работа (например в проекта *MONDILEX,* 7 FP, 2008–2010 – създавания електронен българо-полски речник (Dimitrova et al. 2010: 147–162).

В движението към универсалното DTD описание на речниковата статия, неутрално към конкретен език и конкретен речник, моделът CONCEDE набелязва групи от налична в речниковата статия информация, които се отнасят към структурната йерархия на елементите на речниковата статия, съдържателната им част (значения и дефиниции) и алтернативни решения.

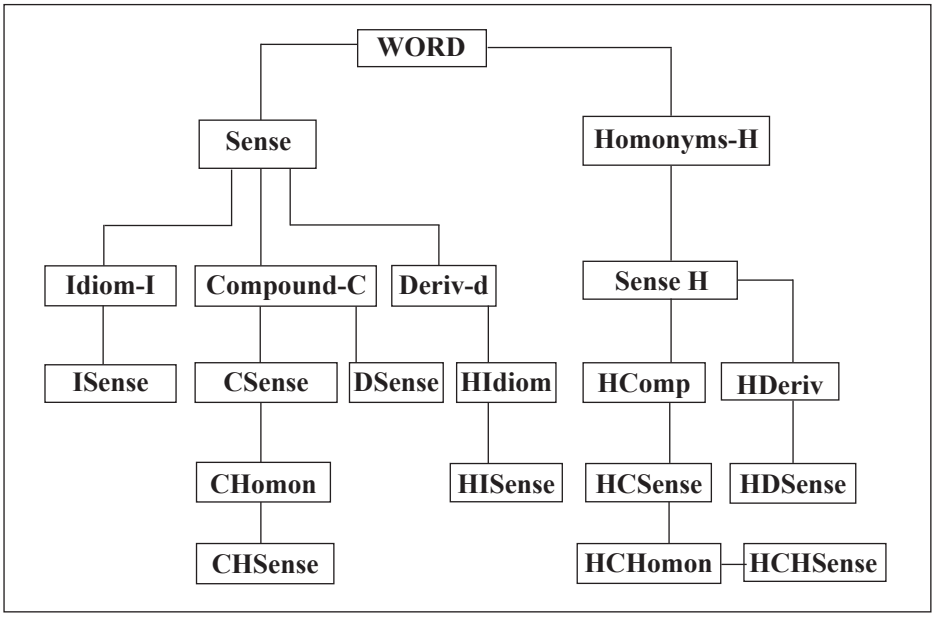
Пътят от тази абстрактна схема към конкретна речникова статия или обратно е доста тежък и в споменатия проект е извървян полуавтоматично за 500 речникови статии.

Реалните нужди на речниковото производство обаче нанасят известни поправки в това идеализирано движение. Тези поправки са свързани преди всичко с отказа от създаване на практически инструмент за реализация на универсален абстрактен речников модел във всякакъв речник – тълковен, двуезичен, за всеки език или двойка езици.

Този отказ е още повече валиден, когато става дума не за създаване на речник от нулата, а за конвертиране на съществуващ печатен речник в електронна форма. Не разглеждаме простата справочна система, където по зададена заглавна дума се получава речниковата ѝ статия – повечето от многобройните електронни речници са от този тип и спестяват само досадното прелистване на хартиени страници. Задачата е да се осъществи преходът от структурирания речников материал към система за управление на база от данни (с възможности за справки по всички възможни комбинации от структурни, йерархични и съдържателни компоненти на речниковата статия).

Такъв преход като конкретно приложение – създаване на електронна версия на съществуващ тълковен речник, бе поставен в изследователски и приложен аспект в Секцията за лингвистично моделиране (в състава на ЦЛПОИ – БАН). В началната си фаза експериментите целяха създаването на дружелюбен интерфейс за електронния преход към АБР – KSBD. Маркировката на речниковата статия в SGML формат, макар и улесняваща формализацията, се оказа недостатъчна за производството на истинска речникова база данни. Тези усилия обаче бяха твърде полезни за едно сравнително проучване на различни речникови структури, описващи една и съща лексикална единица

(например в английски тълковен речник, в англо-руски и англо-български). Резултатите от тези съпоставителни описания показаха разлики в структурирането на речниковата статия, произтичащи от много фактори – обема на речника, принципа на построяване на речниковата статия (степента на гнездова организация), принадлежността на автора на речника към дадена лингвистична школа, особено в прокарването на тънката разлика между многозначност и омография. Описателна и количествена оценка на проявите на тези принципи в различни речници е невъзможна дори само за една заглавна речникова единица, при опит да се проследят структурите в следващите нива на йерархията по най-общите принципи на класификацията, посочена на фиг. 2 по-долу.



Фиг. 2. Йерархия и структура на речниковата статия в максимално разгъната структура – заглавна единица, значение, компаунд (дериват или фраза), идиом, омоним и възможни йерархични вариации и комбинации

Реална електронизация на речниковата статия – опити и успехи

Явно утопичните стремежи да се вместят заглавните единици на повече от един речник в обща схема не могат да доведат до създаването на универсален инструмент за генериране или трансформиране на електронни речникови статии. Очевидно формализацията трябва да се извършва върху конкретен материал – избистрена и детайлизирана концепция за създаването му от нула (top-down) или като наличен печатен речник, който да залегне в основата на електронната речникова база данни.

Такава задача бе поставена, успешно решена и получила практическо решение в рамките на проекта DICO-East, българо-румънско партньорство с Женевския университет като разширение на съществуващия проект Dico-Pro (On-line dictionary consultation for language professionals on intranet) по програмата MLIS – Multilingual Information Society. Продуктите на Dico-Pro (шест речника на Collins за двойките езици английско-френски, английско-немски и френско-немски, Речник на английските идиоми и Френско-швейцарско-немски речник) са обслужвали пряко речниковите нужди на университета, неговия преводачески факултет и библиотеките на Романска Швейцария (Armstrong et al. 2000: 1144–1149).

Задачата на българския участник в проекта бе:

1. Да се намери достатъчно представителен български речник – англо-български или български тълковен, с права за петгодишно ползване в интранет.
2. Да се извърши преформатирането на речника от издателските файлове в SGML запис.
3. Да се организира речниковият SGML запис като база данни с възможности за справки и управление.

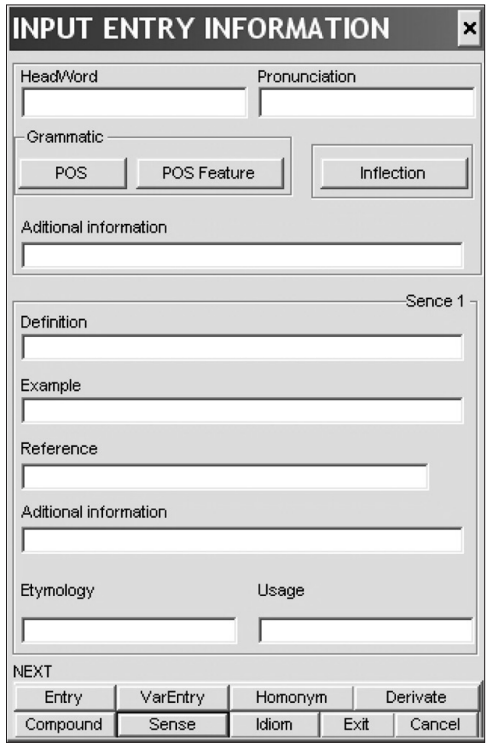
Задачата като практическо приложение бе изпълнена и преизпълнена по параметрите на базата данни – наличните в проекта Dico Pro речници осъществяваха справки само по заглавна дума и плоско текстово търсене в статиите. Вътрешното представяне на речниковата статия като йерархична структура (вж. фиг. 2) в българския речник даваше възможност за по-дълбоки езикови справки на следващ етап.

Тази разработка не бе осъществена на празно място, тъй като две години преди това, в контекста на мултиезиковите европейски изследвания по връзката между корпуси и речник, в рамките на проекта TELRI II бе реализиран програмен продукт за създаване на двуезиков речник и ползването му като база данни. Концепцията за създаване на този продукт, както и интерфейсът на системата DICTATOR (Dictionary Annotation, Compilation and Upgrade), бяха докладвани на IV семинар на проекта „Text corpora and multilingual lexicography“ в Братислава (Paskaleva 1999). Ползването на българския тълковен речник в интранет сървър с търсещи процедури по Dico стандарта бе реализирано на dico-lml.bas.bg за срок от 5 години съгласно договора за авторски права, сключен с притежателя им издателство „Атлантис“. След края на този период речниковата база данни може да бъде ползвана като електронен тълковен речник по други правила на копирайта. По-долу са разгледани решенията на няколко основни проблема на прехода от печатната речникова статия към електронната речникова база данни, улеснени от интерфейса на системата, крайното представяне на речниковата статия в интернетен формат – html, както и възможностите на това представяне за дълбоко изследване на речниковата структура и за производството на нови речници.

Видове входни данни и обработката им

Системата предоставя три начина на обработка на входните данни, в три степени на структуриране. Първият е обикновеното ръчно въвеждане, вторият работи с данни в електронна форма в txt формат, третият – със стандартизиран запис от типа SGML. В зависимост от отговора на въпроса за вида вход, системата се насочва към различна обработка, която извежда експлицитно речниковата структура.

1. Ръчен вход. Трудно можем да си представим ръчно въведен цял речник, но тази опция е полезна за създаване на нови речникови статии. Структурата на въвеждащия интерфейс повтаря структурата от фиг. 2 по-горе, но не в дървовиден, а в последователен вид, с итеративно запълване на повтарящи се части (значения, омографи, вариации от всякакъв род) (вж. фиг. 3).

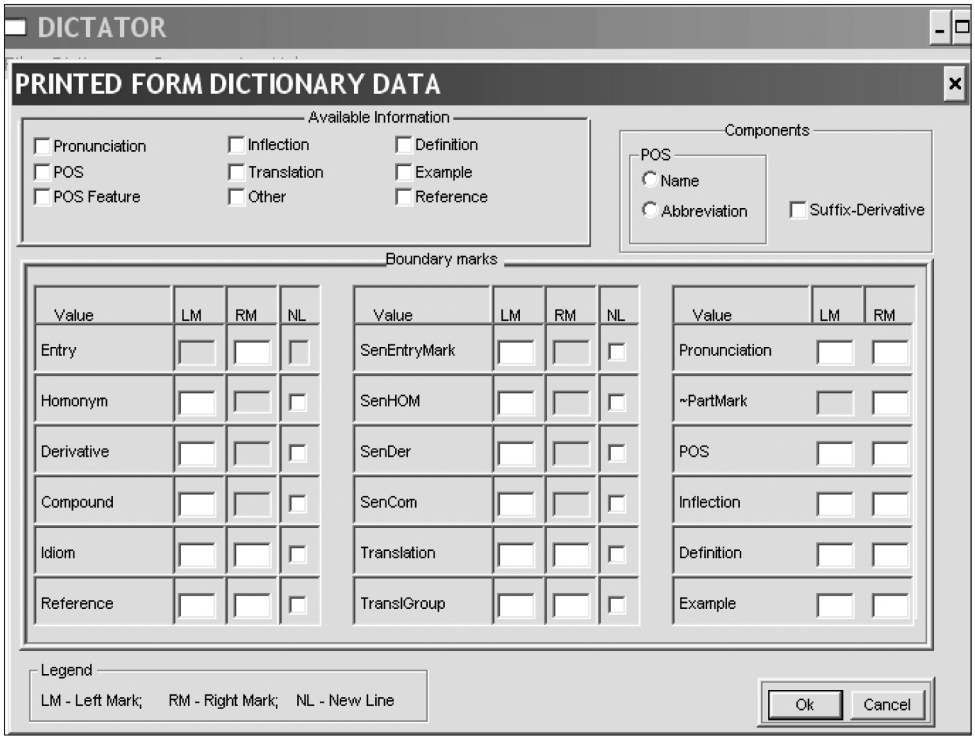


Фиг. 3. Ръчно запълване на речникова статия. Зададени са основните структурни части. Маркерът Next препраща към повтарящите се елементи

2. Генерално (обобщено) запълване на структурата на речниковата статия при наличен електронен запис

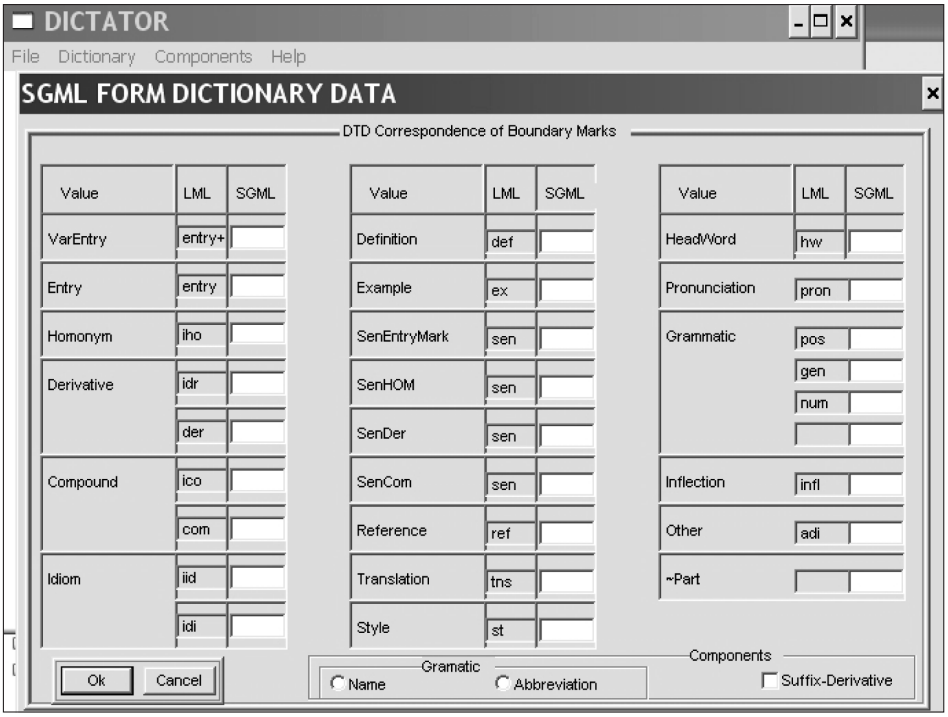
Ако разполагаме с електронен запис на речника в txt формат (независимо от вида на издателските файлове за системата това е единственият

възможен вход), преходът се задава еднократно за всички речникови статии, но затова пък изисква сериозен анализ на формалните признаци за разграничаване на отделните компоненти. Тъй като в текстовия формат липсва информация за шрифт, който има разграничителна функция, значещите маркери се свеждат до ляв/десен съсед на речниковата единица (число, препинателен знак или комбинация от тях), а също и наличие на знак за нов ред. С отчитане и на линейната свързаност на отделните компоненти, при едно грижливо и последователно изчисление е възможно да се определи йерархичната им позиция на всеки компонент при дадени стойностите на горните маркери (вж. фиг. 4).



Фиг. 4. Обобщена форма за запълване на възможните маркери – разграничители на речниковите компоненти

3. Третият начин за въвеждане на речниковите данни бележи най-лесния преход – от стандартизирано SGML представяне като в АБР – KSBD (вж. фиг. 1). Преходът е лесно изчислим, понеже се изискват само еднозначните съответствия между употребените във входната анотация маркери и вътрешните маркери на системата, нужни за по-нататъшното оформление на речниковата статия (вж. фиг. 5).



Фиг. 5. Съответствия между маркерите във вътрешното представяне на речника и входното SGML представяне

Изходно представяне на речниковата статия

Очевидно изходното представяне на така обработените входни речникови данни е самата речникова статия. По стария армейски принцип: „тежко в ученето, леко в боя“ разказаните по-горе изчисления, преформатирания и подредби, веднъж извършени върху целия речников материал, трябва да ни позволят да получим речникова структура във вид на база данни с максимално раздробяване на значещите компоненти и лесно изчислими справки и обобщения за поведението на компонентите от всяко йерархично ниво. Това – що се отнася до вътрешната, изследователската, собствено лексикографската страна на положените усилия. Друга, не по-малко важна страна е приложната, свързана непосредствено с речниковото производство. Системи от изложения тип дават възможност както да се произведе нов речник, така и да се допълват вече произведените не само по обем, но и по структура и богатство на информацията, също и по начин на представяне. Става дума не само за типографското оформяне на речника, но и предоставянето му в различни форми на ползване както в стандартна печатна, така и в електронна форма, а също и в уебпространството.

На такава база – формална изчислимост и електронна обработка с дружелюбен интерфейс, бе реализиран и речниковият сървър на Секцията за лингвистично моделиране (днес в състава на Института за информационни и комуникационни технологии на БАН) от 2004 до 2009 г.

Речниковият сървър предоставя за справки на ограничен кръг клиенти по строги правила на достъп речника ТР – ФС.

Той бе предоставен от издателя в електронна форма (Word for DOS). След преформатиране и изчисляване на структуроопределящата роля на текстовите маркери съобразно с концепцията на речниковата статия стана възможна трансформацията печатен речников запис → речникова база данни (описана по-горе при втория тип входни данни).

Резултатното представяне на речника в html запис възстановява форматирането на статията, като към шрифтовото прибавя и цветово оформление. Шрифтът и цветът имат различителна функция за съответния речников компонент. Например: заглавна дума – червен получер, граматически свойства – зелен курсив, примери – черен курсив, стилистични признаци – светлосин курсив, идиомите – тъмночервен получер и т.н. Вж. фиг. 6.

пара[1] [п`ара] *ж.* **1.** Газообразно, въздухообразно състояние на вода или друга течност. *Водни пари. Алкохолни пари. Водата се превръща в пара.* **2.** *спец.* Газообразно състояние на всяко вещество, което при обикновени условия е течно или твърдо. *Живачни пари.* **Под пара съм** 1. За машина, влак и подобони – готов съм да заработя. 2. *разг.* В напрежение съм, готов съм да започна работа. **С пълна пара** *разг.* - с голямо напрежение, с всички сили. **Вдигам пара** *разг.* – сърдя се, карам се. . **парен** [п`арен] *прил.*

Фиг. 6. Запис от речниковия сървър на омографа *пара*

Това оформление на речниковата статия не е подчинено на естетически критерии. Шрифт, цвят, пунктуация и линейно разполагане, комбинирани в различните си стойности, определят еднозначно мястото на всеки елемент в йерархията на базата данни и улесняват прехода „source html запис – речникова база данни с възможности за търсене и управление“.

Пожелателно заключение

Натрупаният опит и практическите изводи от гореописаната реализация, както и изминатият дълъг път към нея водят до извода, че:

- пълноценното и съвременно високотехнологично използване на продукти като тълковния речник се оказа възможно за кратък период от време с помощта и в рамките на външно финансиране и с неговата логистична и

правна подръжка. Това – при условие, че държим на законите на авторското право както в речниковото, така и в интернетното производство. Този е пътят, по който трябва да се развива високотехнологичният речник, а всичко, което заобикаля този път, е хакерство и интелектуално пиратство.

ЛИТЕРАТУРА

Armstrong et al. 2000: Armstrong, S.., C. Brace, D. Petitpierre, G. Robert and D. Walker. DicoPro: An Online Dictionary Consultation Tool for Language Professionals. // *Proceedings of Euralex 2000*, pp.1144–1149, http://www.issco.unige.ch/en/research/projects/dico.html

Dimitrova et al. 2010: Dimitrova L., V. Koseska-Toszewa, R. Garabik, T. Erjavec, L. Iomdin, V. Shyrokov. Mondilex – Towards the research infrastructure for digital resources in Slavic lexicography. // *Cognitive Studies / Etudes Cognitive Studies*. Vol. 10, pp. 147–162.

Ide and Veronis 1995: Ide, N. and J. Veronis. Encoding Dictionaries. // *Computers and the Humanities* 29, pp. 167–195.

Nerbonne et al. 1997: Nerbonne J., E. Paskaleva, L. Karttunen, G. Proszeky, T. Roosmaa. Reading more in Foreign Languages. // *Proceedings of Fifth Applied Natural Language Processing Conference*. Washington: ACL, pp. 135–138.

Paskaleva 1999: Paskaleva E. The Structure of Bilingual Lexicon Entry Viewed by DICTATOR (A Software Tool for Dictionary Annotation, Compilation and Upgrade). // *TELRI Newsletter No. 9. 4th TELRI European Seminar „Text Corpora and Multilingual Lexicography“*. Bratislava, Slovakia, November 5–7, 1999.

ИЗТОЧНИЦИ

KSBD: *Kernerman semi-bilingual dictionaries.* Kernerman Publishing Ltd. 1986–1993.

АБР – KSBD: *Английско-български речник. Тълковен и двуезичен. Kernerman semi-bilingual dictionaries*. Изд. „Хемус“, 1992. 728 с.

ТР – ФС: Дечева, Д. *Тълковен речник с фразеологически съчетания.* София: Атлантис, 1997. 891 с.